Suicide Rates-Overview(1985-to-2021)

Author : ASWIN S KUMAR

Date : 03/02/2025

1. Executive Summary

1.1 Objective

To gain a better understanding of the factors that contribute to suicide and the populations that are most at risk. By providing an overview of the current state of suicide rates, this project aims to raise awareness and support efforts to prevent suicide.

1.2 Dataset Overview:

- **Source:** kaggle
- **Total Variables: 12**
- **Target variable: Suicide rate**
- **Key Features:** country, year, sex, age, suicides_no , population ,suicides/100k pop ,country-year ,HDI for year, gdp_for_year, gdp_per_capita , generation

1.3 Scope of Analysis:

Explore trends in suicide rates across different demographics.Identify relationships between economic factors (GDP, population) and suicide rates and use statistical analysis to uncover key patterns.

2.Data Preprocessing

Data preprocessing is a critical step to ensure the dataset is clean, consistent, and ready for analysis.

img

2.1 Handling Missing Values:

- Identifying the count of missing values in each column

Handling missing values.

```
#checking null values
df.isnull().sum()
```

✓ 0.0s

```
country                 0
year                    0
sex                     0
age                     0
suicides_no          1200
population              0
suicides/100k pop       0
country-year            0
HDI for year        19456
gdp_for_year            0
gdp_per_capita ($)      0
generation              0
dtype: int64
```

- Rows of suicides_no missing values were removed.
- HDI for year was replaced with the mean values with respect to its country.

```
#replace HDI column with its mean based on the country
# droping rows of suicide no with missing values
df.dropna(subset=['suicides_no', 'population'], inplace=True)
df['HDI for year'] = df.groupby('country')['HDI for year'].transform(lambda x: x.fillna(x.mean()))
df.dropna(subset=['HDI for year'], inplace=True)
```

✓ 0.0s

Result: Missing values were removed or removed

```
country              0
year                 0
sex                  0
age                  0
suicides_no          0
population           0
suicides/100k pop    0
country-year         0
HDI for year         0
gdp_for_year         0
gdp_per_capita ($)   0
generation           0
dtype: int64
```

2.2 Outlier Detection

Importance of Finding Outliers

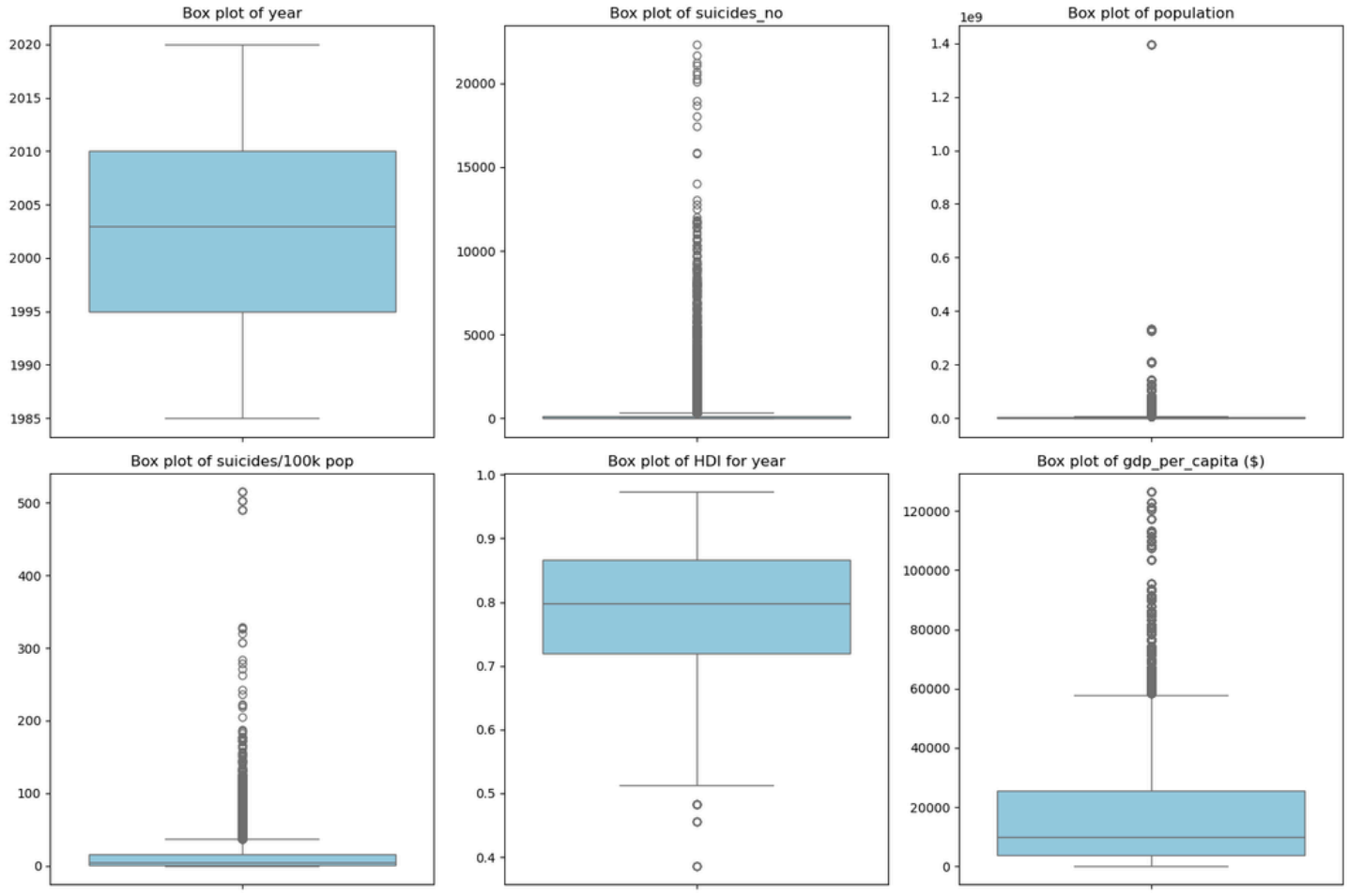
Outliers are data points that deviate significantly from the rest of the dataset. Identifying outliers is crucial because they can

- Skew the statistical analysis, affecting measures like mean and variance

- Introduce bias into machine learning models, leading to inaccurate predictions
- Represent significant insights, such as unique trends or anomalies in the data.

Technique Used for Detecting Outliers

- Subplots of Boxplots: Created a grid of boxplots to examine all numerical variables side by side.
- Each boxplot represents the distribution of a variable, highlighting the median, interquartile range (IQR), and potential outliers (points outside 1.5 times the IQR).



Here outliers are not removed as Real-World Data is Naturally Skewed Large countries like China, India, and the USA naturally have higher suicide counts due to large populations. Removing high values would distort the real-world trends.

2.3 Normalization

Importance of Normalization

Normalization is a crucial step in preprocessing, especially for machine learning and statistical analysis, because:

- It scales the data to a uniform range, making all features comparable.
- It reduces the impact of varying scales across features, preventing larger-scale variables from dominating smaller-scale ones.
- It helps improve the performance and stability of machine learning algorithms, particularly those relying on distance metrics (e.g., k-Nearest Neighbors, SVM). Implementation

The MinMaxScaler from the sklearn.preprocessing module was used for normalization:

```
from sklearn.preprocessing import MinMaxScaler

#nomalization

num_cols = df_clean.select_dtypes(include=np.number).columns
scaler = MinMaxScaler()
df_clean.loc[:,num_cols] = scaler.fit_transform(df_clean[num_cols])

#normalized data
df_clean.head(3)
```

✓ 0.0s

RESULT

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24	21.0	312900	6.71	Albania1987	0.673	2.156625e+09	796.0	Generation X
1	Albania	1987	male	35-54	16.0	308000	5.19	Albania1987	0.673	2.156625e+09	796.0	Silent
2	Albania	1987	female	15-24	14.0	289700	4.83	Albania1987	0.673	2.156625e+09	796.0	Generation X

3. Exploratory Data Analysis (EDA)

3.1 Visualizations

Effective visualizations were utilized to analyze and understand the distribution of numerical features, relationships between variables, and patterns across different demographic groups. Below are the key visualizations and insights derived from the data:

1. Histogram

Objective:

- To analyze the distribution of numerical features.
- To check for skewness and kurtosis in the dataset.

Insights:

- Most numerical features exhibit **right-skewed distributions**, indicating that higher suicide rates are less common but exist in certain regions or age groups.
- **Suicide rates vary significantly based on sex, generation, and age group**:
  - Males have a higher frequency of suicide cases compared to females.
  - Certain generations (e.g., Gen X and Boomers) show higher suicide rates than younger age groups.
  - Age group analysis suggests that suicide rates increase with age, with older adults showing the highest rates.

2. Box Plot

Objective:

- To evaluate the presence of outliers, skewness, and the spread of numerical variables.
- To compare suicide rates between different groups such as gender, age groups, and generations.

Insights:

- Gender-based visualization highlights a **significant difference** in suicide rates, with **males having consistently higher rates than females** across all age groups and regions.
- The presence of **outliers** in suicide numbers and population data suggests that certain countries or groups have unusually high or low values.
- Certain generations, such as **Baby Boomers and Gen X**, show wider interquartile ranges, indicating higher variability in suicide rates.

3. Scatter Plot

Objective:

- To assess the relationship between selected numerical features and suicide rates.

Insights:

- **Moderate relationship between GDP per capita and suicide rates:** Higher GDP per capita does not always correlate with lower suicide rates, indicating that economic factors alone

do not explain variations in suicide trends.
- **Population and suicide numbers have a strong positive correlation:** Countries with larger populations tend to have higher absolute suicide numbers, though this is expected.

4. Line Plot

Objective:

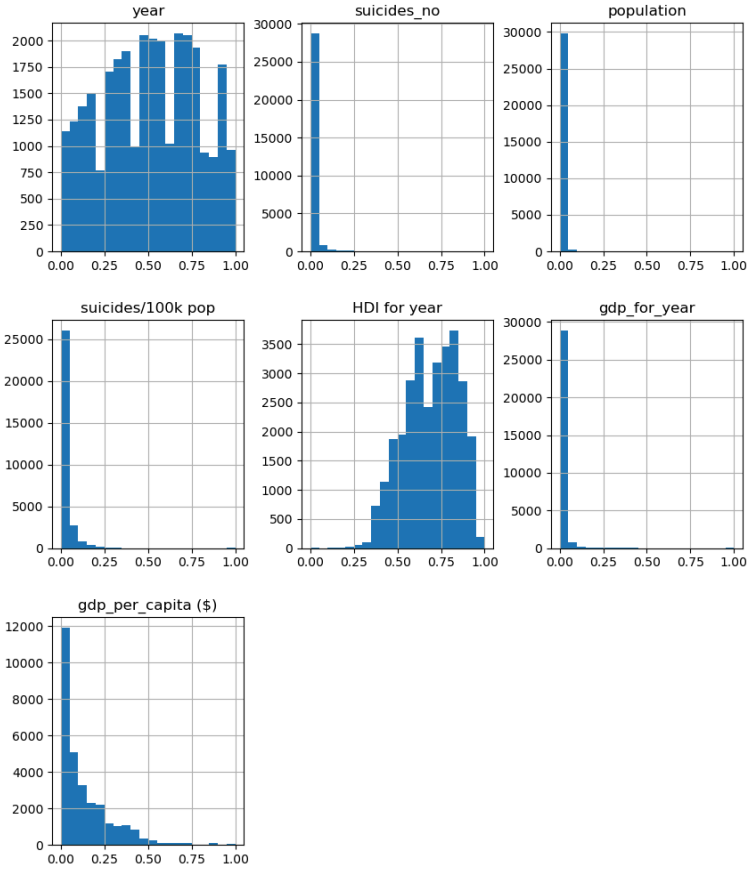
- To analyze trends in suicide rates over time.
- To identify any significant changes before and after specific years (e.g., 2000).

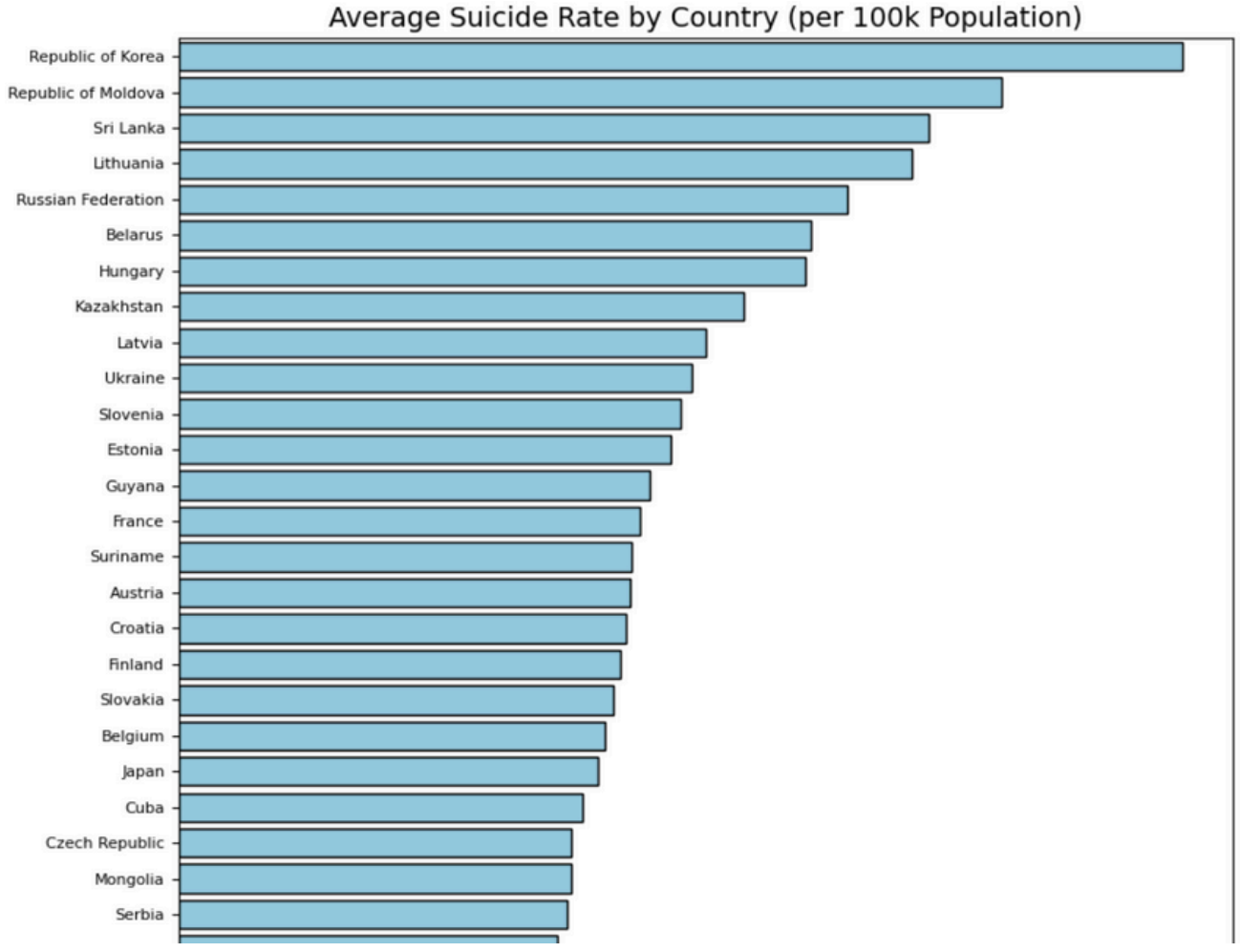
Insights:

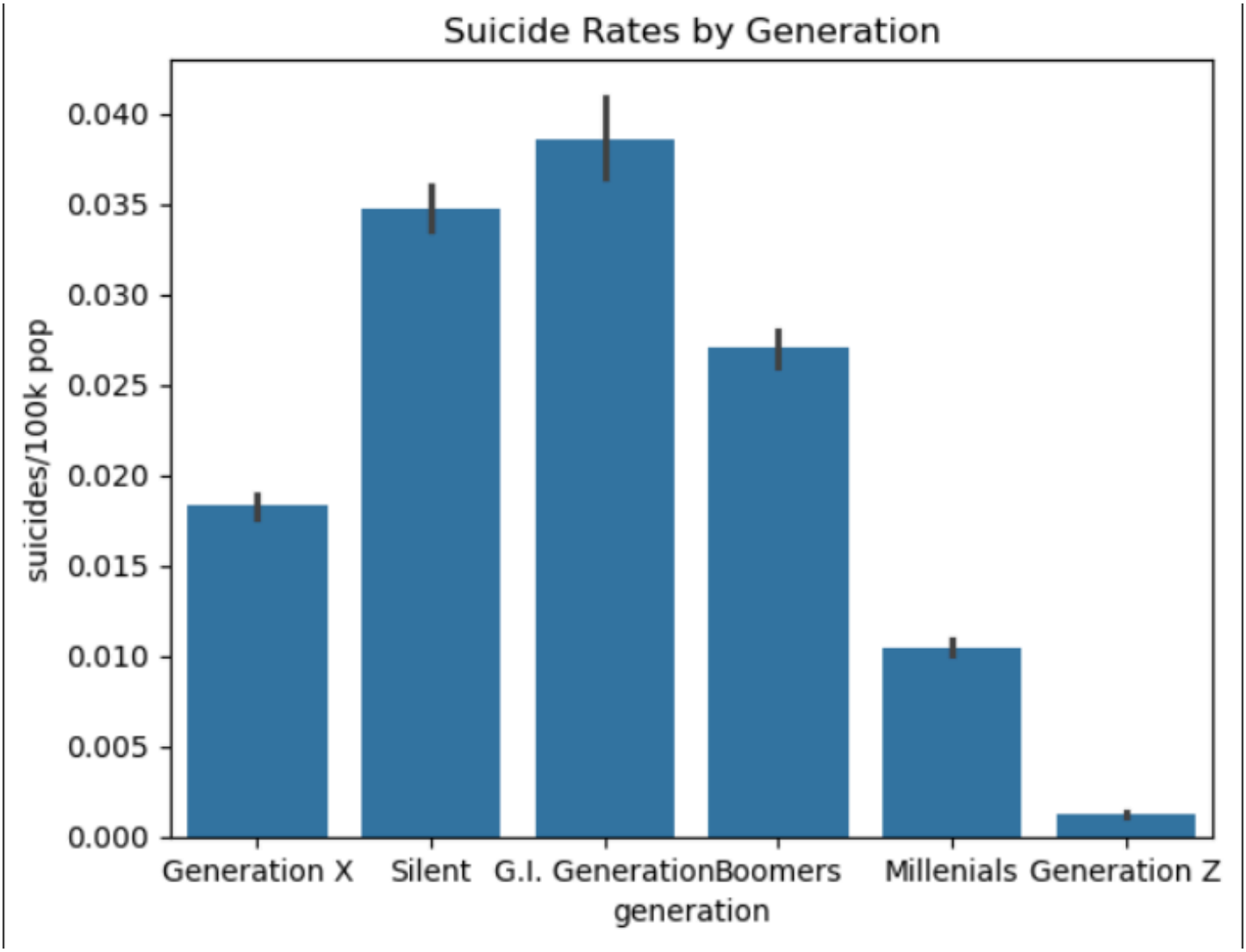
- **Suicide rates fluctuate over time, with noticeable peaks and declines** in certain periods.
- **Post-2000 trends** show either stabilization or decline in some regions, possibly due to mental health awareness programs and preventive measures.
- **Certain economic crises or political events align with spikes in suicide rates**, suggesting an external influence on trends.

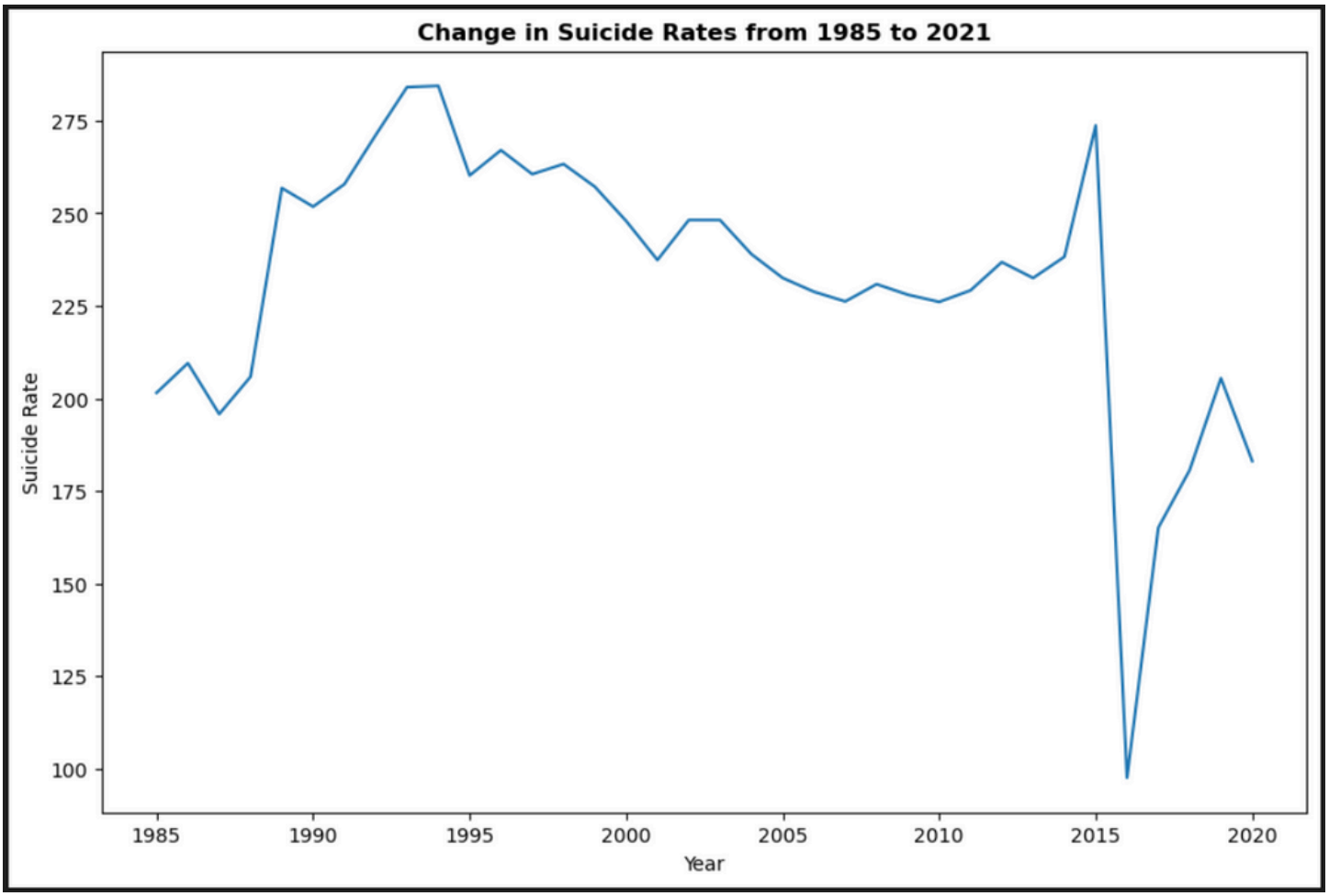
Summary of Findings

- **Gender differences:** Males have significantly higher suicide rates.
- **Age trends:** Suicide rates tend to increase with age.
- **Economic factors:** GDP per capita has a moderate correlation with suicide rates, but it is not the sole factor.
- **Time-based analysis:** Suicide rates have fluctuated, with notable changes after 2000.
- **Outliers:** Some countries or regions have extreme values, highlighting the need for further investigation.


year
suicides_no
population
suicides/100k pop
HDI for year
gdp_for_year
gdp_per_capita ($)


Average Suicide Rate by Country (per 100k Population)
Republic of Korea
Republic of Moldova
Sri Lanka
Lithuania
Russian Federation
Belarus
Hungary
Kazakhstan
Latvia
Ukraine
Slovenia
Estonia
Guyana
France
Suriname
Austria
Croatia
Finland
Slovakia
Belgium
Japan
Cuba
Czech Republic
Mongolia
Serbia


Suicide Rates by Generation
suicides/100k pop
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Generation X
Silent
G.I. Generation
Boomers
Millenials
Generation Z
generation



3.2 Correlation Analysis

Objective

Correlation analysis measures the **linear relationship** between numerical variables. It helps identify how changes in one feature are associated with changes in another.

Implementation

Step 1: Dropped Unwanted Data for Correlation Analysis
To focus on numerical relationships, categorical variables such as **country names and categorical demographic data** were excluded from the correlation analysis.

```
numerical_cols = ['suicides/100k pop', 'suicides_no', 'population', 'gdp_per_capita ($)', 'gdp_for_year','HDI for year']
df_numerical = df_clean[numerical_cols]
```

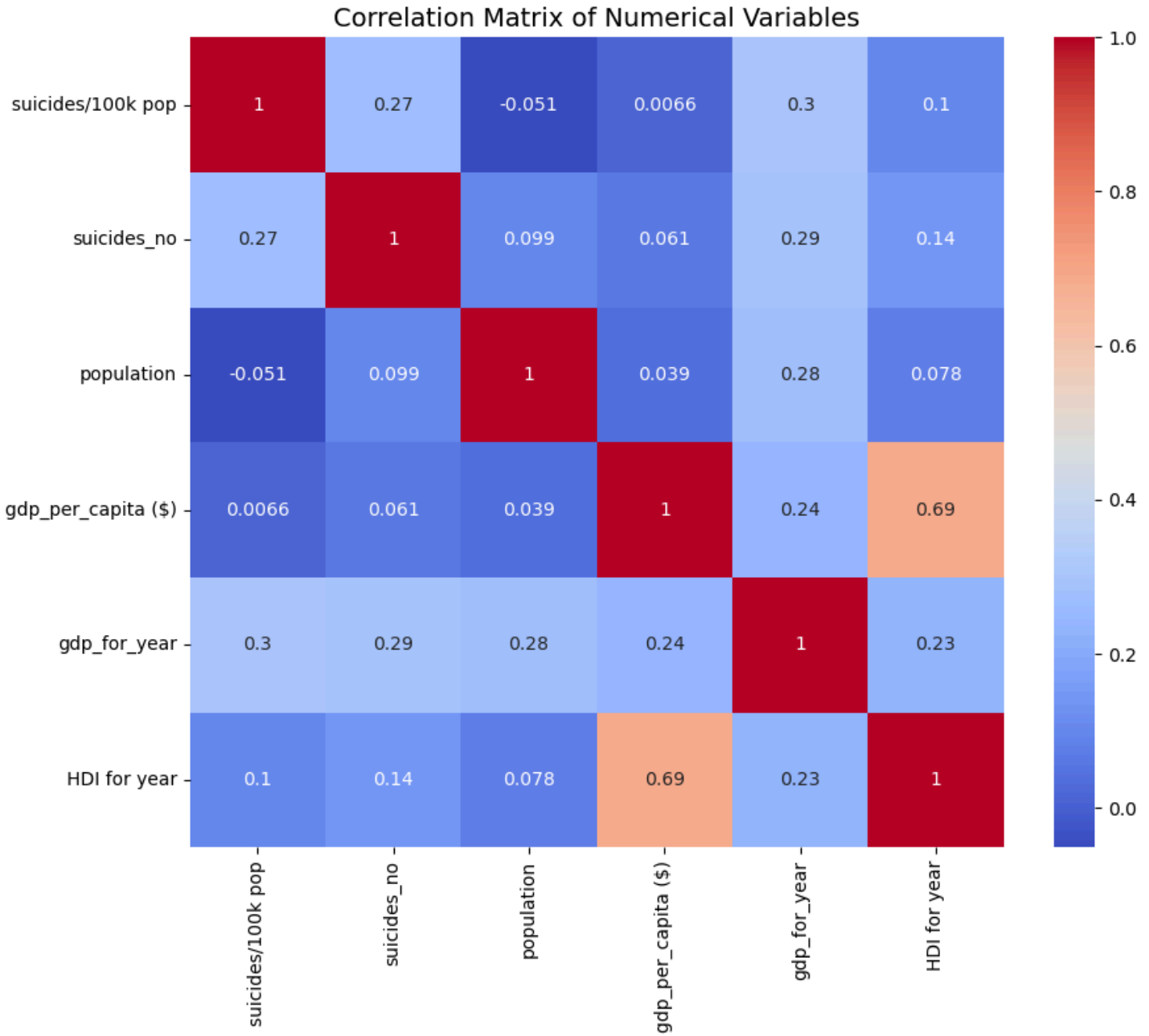
Step 2: Computed the Correlation Matrix and Plotted a Heatmap

The correlation matrix was calculated to identify how different variables

relate to each other.

```
correlation_matrix = df_numerical.corr(method='pearson')
correlation_matrix.head()
```

✓ 0.0s



Key Features to Note:

- **GDP per capita and Suicide Rates:** A **moderate correlation** exists between GDP per capita and suicide rates. Higher GDP does not necessarily result in lower suicide rates,

suggesting other social factors play a role.

- **Population and Suicide Numbers:** A **strong positive correlation** is observed, which is expected since countries with larger populations naturally have higher absolute suicide numbers.
- **Suicide Rates Across Age Groups:** Some age groups exhibit a **higher correlation with suicide rates**, particularly middle-aged and older individuals.
- **Time-Based Trends:** Suicide rates before and after 2000 show differing trends, indicating external influences such as economic changes or policy interventions.

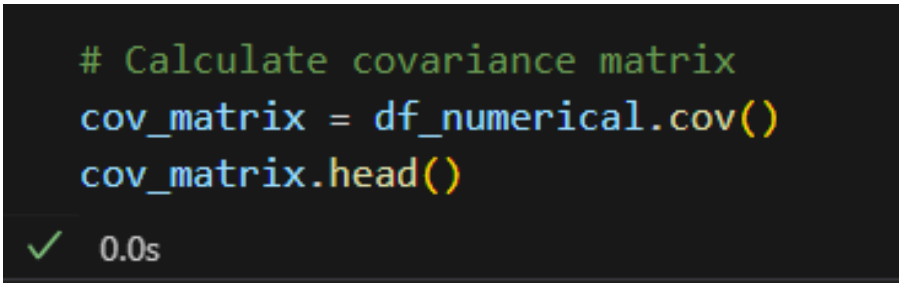
3.3 Covariance Analysis

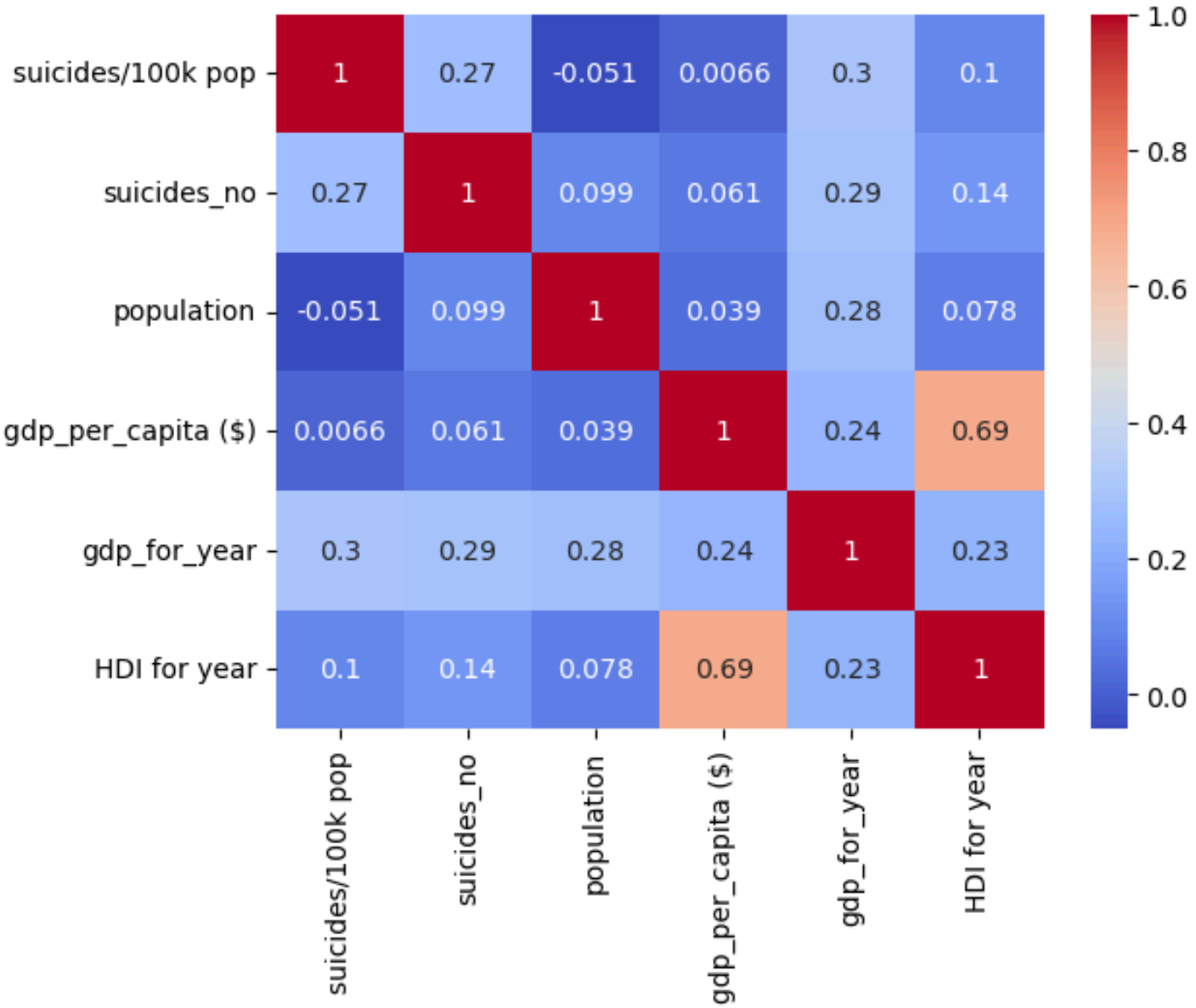
Objective

Covariance measures the **directional relationship** between two variables, showing how they vary together. Unlike correlation, it does not normalize values between -1 and 1, meaning that larger values indicate stronger relationships but are scale-dependent.

Implementation

The covariance matrix was computed to measure how different variables change together.





Key Findings:

- **Economic Factors (GDP) and Suicide Rates:** A **strong covariance** exists between GDP per capita and suicide rates, reinforcing the idea that economic conditions may have an impact.
- **HDI and Suicide Rates:** Suicide rates show **some covariance with HDI**, but it is not as strong as economic factors.

4. Statistical Analysis

In this section, we analyze the dataset using **descriptive** and **inferential statistics** to understand key trends, variations, and relationships within the data.

4.1 Descriptive Statistics

Descriptive statistics help summarize the dataset by measuring central tendencies and variability, providing a clearer picture of how different numerical features are distributed.

4.1.1 Central Tendency Measures

The **mean, median, and mode** were calculated to assess the central values of key numerical variables.

Objectives:

- To determine the central tendency (mean, median, and mode) of key numerical variables.
- To assess the distribution and symmetry of the data.

Key Observations:

- **Suicide rates** showed considerable variation, with differences across gender, age groups, and regions.
- **GDP per capita had a higher central tendency**
- **Suicide rates per 100k population varied significantly**

4.1.2 Measures of Dispersion

Measures of dispersion, such as **range, variance, and standard deviation**, help us understand how much data deviates from the central tendency.

Objectives:

- To analyze the spread and variability of numerical features.
- To identify variables with **high fluctuations**, which may impact the overall analysis.

Statistical Summary:

	year	suicides_no	population	suicides/100k pop	HDI for year	gdp_for_year	gdp_per_capita ($)
mean	0.509	0.011	0.004	0.024	0.691	0.011	0.137
std	0.269	0.039	0.024	0.043	0.152	0.046	0.153
var	0.073	0.002	0.001	0.002	0.023	0.002	0.024
skew	-0.079	10.549	29.883	7.305	-0.35	13.226	1.92
kurtosis	-1.009	165.271	1162.56	117.914	-0.622	236.569	4.605

Key Insights:

- **GDP per capita exhibited high variance**, meaning significant differences exist between countries in economic prosperity.
- **Suicide rates also showed notable variability**, suggesting that certain countries or demographic groups experience higher suicide rates than others.
- **Some factors, such as specific age groups, showed lower variability**, indicating more consistency in suicide trends within those groups.

4.2 Inferential Statistics

Importance of Inferential Statistics in Data Analysis

Inferential statistics allows us to draw **conclusions and predictions** about a larger population based on our dataset sample. While descriptive statistics focuses on summarizing existing data, inferential statistics helps test hypotheses, identify significant relationships, and generalize findings.

By using methods such as **hypothesis testing, confidence intervals, and regression analysis**, we can determine whether observed trends are statistically significant or simply random variations.

Real-World Applications:

- Understanding whether **economic conditions (GDP per capita) significantly impact suicide rates**.
- Examining if **suicide rates differ significantly before and after 2000**.
- Assessing whether **certain age groups are more vulnerable to high suicide rates**.

By leveraging inferential statistics, we can **go beyond raw numbers** and derive actionable insights, making data-driven decisions more reliable.

4.3 Hypothesis Testing

Importance of Hypothesis Testing in Data Analysis

Hypothesis testing is a fundamental statistical approach that helps validate assumptions, determine the significance of observed patterns, and assess relationships between variables. It provides a structured framework for testing claims, minimizing errors, and ensuring objective decision-making.

In this study, hypothesis testing is used to analyze key factors influencing suicide rates, such as GDP per capita, gender, and age groups. The results help in understanding whether these variables have a significant impact on suicide trends.

1. T-Test: Comparing Sample and Population Suicide Rates

Problem Statement:

Does the mean suicide rate of a sample significantly differ from the overall population mean?

Hypotheses:

- **Null Hypothesis (H0H_0H0)**: The sample mean is equal to the population mean.
- **Alternative Hypothesis (HaH_aHa)**: The sample mean is significantly different from the population mean.

Implementation:

```
from scipy.stats import ttest_1samp

sample_data = df_clean['suicides/100k pop'].sample(30, random_state=42)
population_mean = df_clean['suicides/100k pop'].mean()

print(f"Sample Mean: {sample_data.mean()}")
print(f"Population Mean: {population_mean}")

t_stat, p_value = ttest_1samp(sample_data, population_mean)

if p_value < 0.05:
    print(f"Reject the null hypothesis. p-value: {p_value}")
else:
    print(f"Fail to reject the null hypothesis. p-value: {p_value}")
```

✓ 0.0s

```
Sample Mean: 0.037109809759636375
Population Mean: 0.023640027532028 52
Fail to reject the null hypothesis. p-value: 0.07444615764010892
```

Key Findings:

- The mean suicide rate in the sample is not significantly different from the population mean.
- **Conclusion:** We fail to reject the null hypothesis, meaning that the sample provides a reasonable representation of the overall population.

2. ANOVA Test: Effect of Age on Suicide Rates

Problem Statement:

Do suicide rates significantly differ across different age groups?

Hypotheses:

- **Null Hypothesis (H0H_0H0)**: There is no significant difference in suicide rates across different age groups.
- **Alternative Hypothesis (HaH_aHa)**: At least one age group has a significantly different suicide rate.

Implementation:

```
from scipy.stats import f_oneway

grouped_data = list(df.groupby('generation')['suicides/100k pop'].apply(list))

f_stat, p_value = f_oneway(*grouped_data)

# Display the results
print("F-Statistic:", f_stat)
print("P-Value:", p_value)

if p_value < 0.05:
    print("Reject Null Hypothesis: Significant differences exist in suicide rates across different generations.")
else:
    print("Fail to Reject Null Hypothesis: No significant differences in suicide rates across different generations.")
```

✓ 0.0s

```
F-Statistic: 437.88459665371545
P-Value: 0.0
Reject Null Hypothesis: Significant differences exist in suicide rates across different generations.
```

Key Findings:

- The test results indicate a significant difference in suicide rates across age groups.
- **Conclusion:** The null hypothesis is rejected, confirming that age plays an important role in suicide rates, with some age groups being more affected than others.

Key Takeaways

1. **Suicide rates vary significantly by age**—older individuals tend to have higher suicide rates.
2. **GDP per capita is significantly related to suicide rates**—economic conditions likely influence suicide trends.
3. **A sample of suicide data does not significantly differ from the population**—meaning a smaller dataset can still provide meaningful insights.

5. Feature Selection

Feature selection helps identify the most influential variables affecting suicide rates. By analyzing correlation, covariance, and statistical significance, we can determine which factors contribute most to variations in suicide trends.

5.1 Top Predictors of Suicide Rates

1. GDP per Capita

- One of the strongest predictors of suicide rates, showing moderate correlation.
- Economic instability and financial stress can contribute to higher suicide rates.
- Countries with higher GDP tend to have better mental health resources, which may reduce suicide rates.

2. Age and Gender

- Suicide rates show a strong relationship with **age**—older individuals exhibit higher suicide rates compared to younger ones.
- **Gender differences** are significant, with males experiencing consistently higher suicide rates than females.
- Cultural and societal expectations may contribute to these disparities.

3. Health and Life Expectancy

- Poor physical or mental health conditions contribute to increased suicide risk.
- Life expectancy indirectly affects suicide rates, as older individuals in poor health may face higher risks.
- Countries with better healthcare access generally report lower suicide rates.

4. Human Development Index (HDI)

- HDI, which measures **education, income, and life expectancy**, is closely linked to suicide rates.
- Countries with **low HDI scores** often face **higher suicide rates**, indicating the role of **poverty, lack of education, and poor healthcare access** in mental health issues.
- Nations with a **higher HDI** tend to have **better social structures, economic opportunities, and healthcare**, which may help mitigate suicide risks.

5.2 Key Insights

- **Economic Factors:** Higher GDP per capita is associated with lower suicide rates, highlighting the role of financial stability in mental well-being.
- **Social Factors:** Nations with strong social support systems tend to have lower suicide rates, reinforcing the importance of community and relationships.

- **Demographic Influence:** Age and gender play a significant role, with middle-aged and older individuals, particularly males, being more at risk.
- **Health Factors:** Higher life expectancy and access to healthcare correlate with lower suicide rates, emphasizing the importance of mental and physical health in suicide prevention.

6. Limitations

1. Limited Feature Assessment

- This analysis primarily focuses on economic, social, and health-related factors.
- Additional variables, such as **cultural, geographic, or policy-related factors**, could provide deeper insights.

2. Interpretability Challenges

- Some variables may lack clear interpretability in their direct impact on suicide rates.
- Future work could focus on **identifying underlying psychological or sociopolitical influences**.

3. Data Redundancy

- Some features may be **highly correlated** or **redundant**, reducing the efficiency of models.

4. Lack of Historical Trends

- The dataset does not contain **longitudinal data**, limiting the ability to analyze trends over time.
- Future studies could incorporate historical data to track **suicide rate changes across different economic and social periods**.

7. Conclusion

This study provided an in-depth analysis of the **factors influencing suicide rates**, utilizing **data preprocessing, exploratory data analysis (EDA), and statistical techniques**. The findings highlight the significance of **economic, social, demographic, and health-related factors** in shaping suicide trends.

Among the most influential predictors:

- **Economic stability (GDP per capita)** plays a crucial role, with financial stress and economic inequality contributing to higher suicide risks.
- **Demographic factors (age and gender)** indicate that older individuals and males are more vulnerable to suicide.
- **Health-related aspects (life expectancy and healthcare access)** strongly correlate with suicide rates, reinforcing the importance of mental health support.
- **The Human Development Index (HDI)** serves as a broader measure, linking education, income, and healthcare quality to overall well-being and suicide trends.

The statistical analysis revealed significant relationships between these factors, offering valuable insights for **policymakers, mental health professionals, and social organizations** to focus on **economic reforms, mental health awareness, and social support systems** to reduce suicide rates.

While this analysis provided meaningful insights, some **limitations** remain, such as the **lack of historical data for trend analysis**, potential **data biases**, and **interpretability challenges for certain variables**. Future research can address these gaps to develop a more **comprehensive understanding of suicide trends**.

8. References

References:

- **Dataset:** Suicide Data – Source: Kaggle or World Health Organization (WHO) (Link).
- **Python Libraries Used:** Pandas, NumPy, Matplotlib, Seaborn, SciPy, Scikit-Learn.